Modeling and measuring training information in a network

SML 2014

Jan Ramon

KATHOLIEKE UNIVERSITEIT LEUVEN

Contents

- Modeling example dependencies
- Problem of measuring training information
- Contributions
- Conclusion

Networked examples

- Relationship
  - in which several objects participate
  - with features of these objects
  - with target value depending on all objects
- Examples can share objects
  - Not independently drawn (i.i.d.)

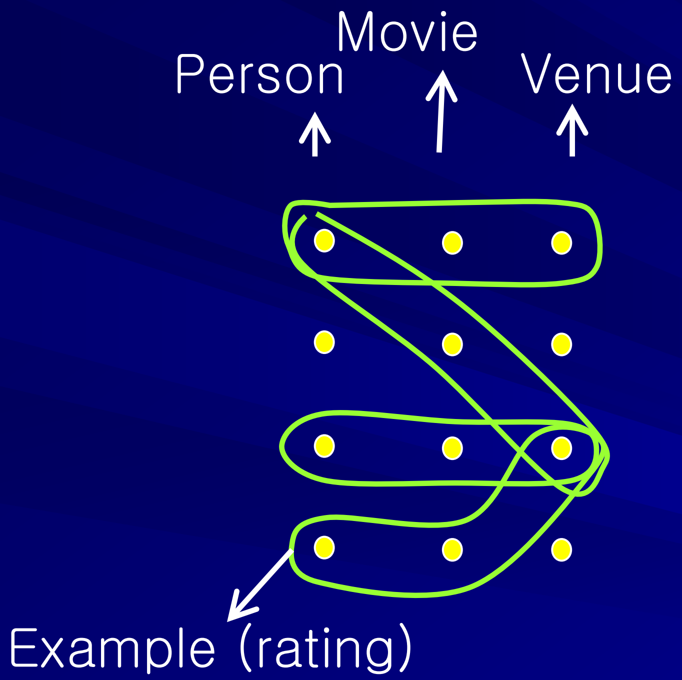
Networked data
Movie rating example

- Movie rating
  - Obj: Movie (genre, duration, actor popularity)
  - Obj: Person (age, gender, ...)
  - Obj: Screening (location, time, ...)
  - Target: Rating
- Training set: several ratings per person / movie / cinema
- Test set: new person / movie / cinema

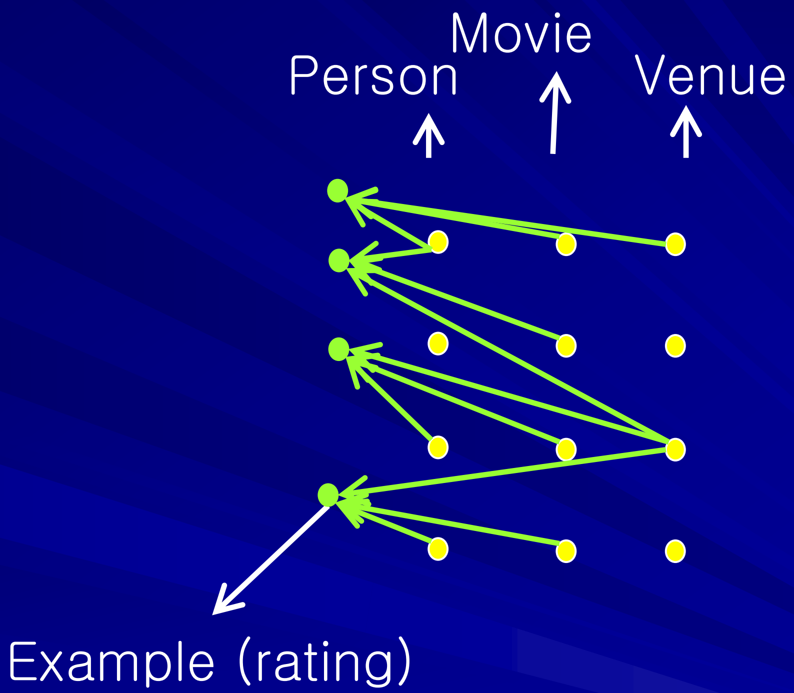
Networked data
Lawsuit example

- Lawsuits:
  - Obj: Person
  - Obj: Case
  - Obj: Judge
  - Target: outcome
- Judges handle several cases, persons may be involved in several cases

Representation



Network based representation



Bayesian network representation

Contents

- Modeling example dependencies
- Problem of measuring training information
- Contributions
- Conclusion

Problem

Practical learning problem

- Given:
  - Network with objects connected by examples
  - Target values
  - Test set with new objects & examples
- Predict:
  - Target values of test examples

Theory: Effective sample size

- How large an i.i.d. dataset should be to contain the same amount of training information?

Related problems (1)

- Link prediction, vertex/edge property prediction ...
  - Predict rating of a person for a movie *based on movies he has seen already* → we should have earlier ratings!

- Not considered here
- Combination may be possible

Related problems (2)

- (Wang, Neville, Gallagher, Eliassi-Rad, ECML/PKDD-2011) :
  - vertices are examples
  - edges indicate a bounded covariance
- safe correction for statistical significance tests
- Both effective sample size (here) and significance (Neville et al.) are potential applications of both models.

Contents

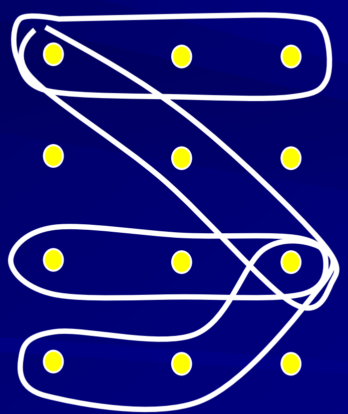
- Modeling example dependencies
- Problem of measuring training information
- Contributions
- Conclusion

Independence assumptions

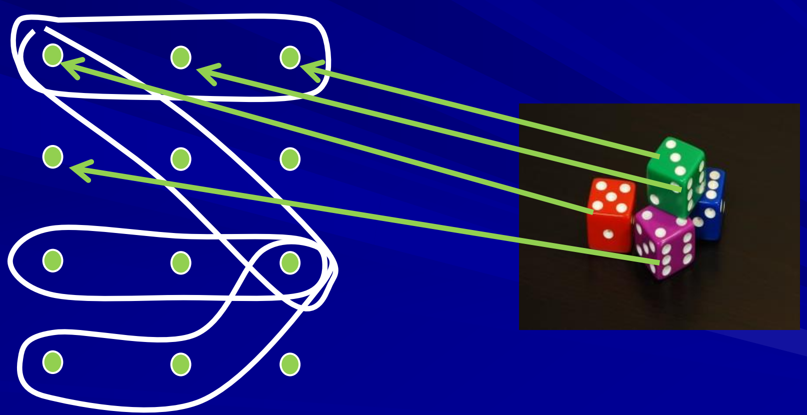
- Weaker form of i.i.d
- But not arbitrary
  - arbitrary $\Rightarrow$ no bound possible

Independence assumptions

- Edges are fixed.
- The features of every vertex are drawn i.i.d. (not even depending on the edges).



1. Choose edges (possibly very dependently)



2. Draw vertex features (don't look at edges)

Independence assumptions applied

YES

- Sneak preview
- Randomized trial: patients are assigned randomly to set of treatment params
- Cases are assigned randomly to judges

NO

- Select movie based on genre
- Patients go to closeby hospital or to hospital recommended by their friends
- Judges handle cases connected to their existing cases

Measuring guaranteed accuracy

- Basic problem: evaluate accurately expected value of a statistic
  - Set of networked examples E
  - Distribution over features D
  - function f on examples
- Determine $\hat{\mu}_{minvar}$ minimizing
$$\mathrm{var}(\hat{\mu}_{minvar} - E_{e\sim D}[f(e)])$$
- For some ϵ, determine $\hat{\mu}_{PAC}$ minimizing
$$1 - \delta = \mathrm{P}(|\hat{\mu}_{PAC} - E_{e\sim D}[f(e)]| > \epsilon)$$

PAC structure

- PAC: with probability $1-\delta$ the loss is bounded by ε where
$$\delta \propto exp\left(\frac{-m(S)\epsilon^2}{C_1 + C_2\epsilon}\right)$$
- with $m(S)$ the effective sample size of training set S. Higer $m(S)$ = better
- i.i.d. sample S, $m(S) = |S|$ = best possible

Effective sample size measures

$$\max(\boldsymbol{n_{EQW}}, \boldsymbol{n_{IND}}) \le n_{MIS} \le \vartheta \le \boldsymbol{s} \le n_{MSC}$$

All weights equal

$n_{EQW} = \frac{|S|}{\chi^*(G)}$ (=2)

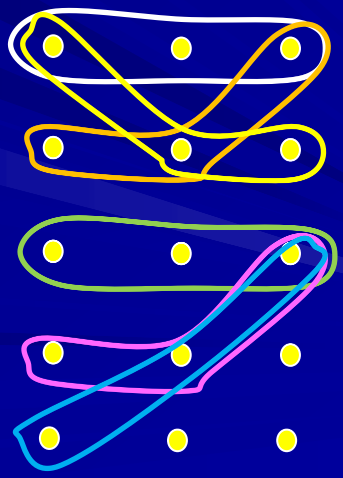
Lovasz measure (=2.5)

Max set cover (=3)

(Max) independent set

$n_{IND} = |IND|$ (=2)

Relaxed (weighted) Max independent set:: LP (=2.5)



Relaxed maximum independent (edge) set

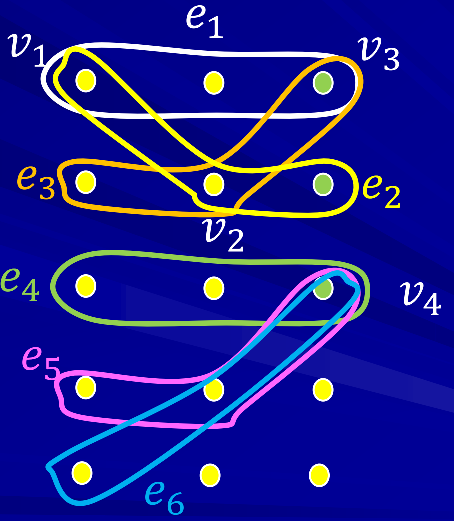
- Influence of each factor (object) is at most 1:

max s

$$s = w_1 + w_2 + w_3 + w_4 + w_5 + w_6$$

s.t.

$$
\begin{aligned}
v_1: &\quad w_1 + w_2 \le 1 \\
v_3: &\quad w_1 + w_3 \le 1 \\
v_2: &\quad w_2 + w_3 \le 1 \\
v_4: &\quad w_4 + w_5 + w_6 \le 1
\end{aligned}
$$



$s = 2.5$

Does it matter?
A theoretical experiment

- BA-graph: $f(d) \sim d^{-3}$

$$\frac{var(\mu_{EQW})}{var(\mu_s)} = \ln(|V|)$$

(for $|V| = 10^7$ → 8.06)

- ER-graph: Δ = avg degree

$$\frac{var(\mu_{EQW})}{var(\mu_s)} = 1 + \frac{1}{\Delta}$$

($\Delta = 4$ → 1.25)

Conclusions

- Learning from networked examples
  - Explicitly modeling dependencies
  - PAC-style bound / effective sample size
  - Computing statistics to minimize variance
- Future work:
  - More accurately modelling dependencies
  - Weaker independence assumptions

Questions?

?